DOCUMENT RESUME

ED 243 321 FL 014 344

AUTHOR Friedman, Charles B.
TITLE The Construct Validation of Language Tests Using Bias Techniques.
PUB DATE Mar 84
NOTE 17p.; Paper presented at the Annual Meeting of the Teachers of English to Speakers of Other Languages (18th, Houston, TX, March 6-11, 1984).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Comparative Analysis; English (Second Language); *Language Tests; *Research Methodology; *Second Language Instruction; *Statistical Analysis; *Test Bias; *Test Validity
IDENTIFIERS *Test of English as a Foreign Language

ABSTRACT

A method for asessing second language test validity by modifying and applying statistical test bias techniques to the performance of language learners from different native language groups was developed out of concern for internal test bias, both overall and item-related. Valid tests should contain some items testing differential performance between different native language groups, but global test measurement should also, theoretically, be similar, providing similar measurement of the same underlying constructs with the same degree of accuracy. That is, micro-level sensitivity and macro-level similarity should exist at the same time. In a pilot study of the Test of English as a Foreign Language (TOEFL), results of testing 481 native Chinese speakers and 484 native Arabic speakers were analyzed by the following methods. To investigate macro-level similarities, the Kuder-Richardson-20 statistical method and a formula suggested by Feldt were used, and a separate factor analysis was performed for each language group for each test section using the SSPS program factor PA2. At the micro level, three analyses were used: two item difficulty parameters (transformed item difficulty and Rasch difficulty) and a chi-square analysis of distribution of correct responses across difficulty levels. Results showed that the TOEFL sections do measure similar constructs with the same degree of accuracy, and that individual items are sensitive to native language difference, giving evidence of the construct validity of the test. (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED243321

THE CONSTRUCT VALIDATION OF LANGUAGE TESTS USING BIAS TECHNIQUES*

Charles B. Friedman
Pan American University
Edinburg, Texas

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Charles B. Friedman

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

FL014344

*Paper presented at TESOL '84 Houston, Texas, March 6-11, 1984

So much research in second language acquisition is product oriented that it is crucial that the manner in which we measure our product indeed truly measure what we think we are measuring. In other words, if we say an instrument measures language proficiency, we must be sure that it in fact does measure that trait. All too often an instrument is used to evaluate either an experiment or student progress without first verifying or validating the instrument we are using. Too often it is assumed that because an instrument was used before by someone, it is a good, valid instrument. Too often it seems we rely on sage advice or test title designation to conclude that a test is measuring what it says it is measuring. All too often, we are being negligent and in some cases unknowlingly committing acts of language testing malpractice. This malpractice can be viewed in terms of the test characteristics themselves or how the test is applied. It is the purpose of this paper to look at the former and suggest one additional perspective in test validation.

All of the preceding comes under the rubric of construct validation. i.e. is the test measuring what we think it is measuring? While language testing has long recognized the central importance and difficulty in validating a test and the central role of theory in both language and test description, it has long neglected the specific area of construct validity. Construct validity, first explicated by Cronbach and Meehl (1955),

is crucial whenever there is an attribute or trait not directly observable eg. language proficiency. More than any other type of validity it relies on the building up of evidence from numerous studies and the application of a variety of techniques to assess the degree to which the test is measuring what it purports to measure. The choice of these techniques rely in part on the ingenuity of the individual researcher.

While in 1974 Stevenson characterized language testing for its lack of concern with construct validity, recent efforts (Bachman and Palmer, 1981, Oller and Hinofotis, 1980) have begun to examine construct validity in more depth. Bachman and Palmer, for example, have made extensive use of the 1959 Campbell Fiske multitrait-multimethod matrix in a number of studies. More work however, remains to be done, and it is important that we not place too much emphasis on one technique.

It is the purpose of this paper to suggest an additional method for second language test construct validation by modifying and applying statistical test bias techniques to the performance of second language learners from different native language groups. Before we begin to examine the specific methodology, a theoretical perspective must be established.

Both external and internal theories of test bias rest on statistical assumptions. External bias deals with the ability of the test to predict a criteria for different groups and usually uses some type of regression analysis. Stevenson (1974) has pointed out the lack of a clear reliable valid criterion for language proficiency. The other concern of bias and of

interest here is internal bias and its application to construct validity.

The first aspect of internal bias applies to the hypothetical traits or constructs being measured by the test. Bias is said to exist in regard to construct validity if the same test measures different traits for different groups or measures those traits with different degrees of accuracy or reliability. (Reynolds, 1982). Thus, whatever underlies the performance of one group's test performance should also underlie another group's performance and do so with the same degree of accuracy. The second concern of statistical bias operates at the micro or item level. An item is viewed as biased if equally able individuals from different groups do not have the same probability of success (Shepherd, et.al. 1981). In summary, traditional statistical views of bias view any discrepant behavior across groups, be it at the macro (whole test) level or micro (item) level, as an indicator of bias.

Numerous studies, for example, have examined the performance of sex and racial groups on intelligence tests. (Jensen 1980) If discrepant systematic behavior for individual items or on the whole test was found, it was then said that the test/item exhibited some level of bias. This measure of bias was viewed as one of degree rather an all or nothing concept.

Strict psychometric application of the above to second language testing and the performance of different native language groups, however, would be a gross error and could lead to inappropriate interpretations. Normally, any evidence of systematic

bias, be it at the macro or micro level, would be evidence of bias and poor construct validity. In language testing and learning one must recognize the unique nature of language and the existence of some influence of native language on learning the target language. One need not subscribe to the strong contrastive analysis hypothesis to accept some native language interference as a source for error and influence in acquisition.

The above suggests a dual approach to internal test bias and second language test construct validity. Evidence of construct validity should be indicated by the existence of "ironic bias". On one hand a number of items should exhibit some differential performance (bias) between different native language groups. This would indicate that the test is sensitive to native language. On the other hand global test measurement should theoretically be similar: at the macro level the test should provide similar measurement of the same underlying constructs with the same degree of accuracy. Theoretically, the evidence of construct validity would be shown by the juxtaposition of micro sensitivity (differences) and macro similarity.

METHOD

To apply the above theoretical perspective, a methodology was devised and applied to a 1979 international administration of the Test of English as a Foreign Language. This sample was restricted to 481 native speakers of Chinese and 484 native speakers of Arabic, a subset of the population used in a study

by Alderman and Holland (1981). Neither language is Indo-European and neither has a Latin writing system. In addition, Arabic is an alphabetic system whereas Chinese is a logographic one. These gross reasons make both Arabic and Chinese distinctly different from English and distinctly different from each other.

In order to evaluate macro level similarities, two methods were used. The first technique tested the null hypothesis that the internal test reliability for Arabic speakers is not significantly different from the internal test reliability for Chinese speakers on each TOEFL test section. Internal test reliabilities were calculated using Kuder-Richardson-20 and then significant differences tested using a formula suggested by Feldt (1969). Finding significant differences and rejecting the null hypothesis would indicate that the test section is not measuring the groups with the same degree of accuracy.

The second technique used in evaluating the macro structure examined the underlying factor structure for each test section. Construct validity would be indicated if the underlying factor loadings for each factor on each test section between the Arabic and Chinese groups were highly correlated. A separate factor analysis was performed for each language group for each test section using SPSS program factor PA2. The number of factors was preset to four with an orthogonal varimax rotation. The usual measure of similarity between groups is the Pearson r; however, in comparing factor loadings, assumptions of normality and linearity are likely to be violated. Therefore, factor

loadings were compared using a coefficient of congruence as suggested by Harman (1976). This coefficient ranges from +1 to -1, similar to Pearson r. Separate SPSS programs were written to calculate the coefficient of congruence for all corresponding factors. Thus, at the macro level of analysis two techniques, one examining degree of accuracy and the other the similarity between corresponding factors structures, were used to evaluate construct validity. Following this, individual item behavior needed to be analyzed and evaluated for sensitivity to the different native language groups.

While various techniques to measure differential item performance at the micro or item level have been proposed, they have often yielded different results (Ironson and Suboviak, 1979; Shepherd, Camilli, and Averill, 1981). These methods differ in their conceptualization of bias, theoretical base, statistical complexity, and parameters used to indicate bias. In order to obtain a more reliable index of item bias and verify that similar items were being detected as sensitive, three different detection methods were employed. Two of the techniques employed an item difficulty parameter, though one is based on a classical testing model (transformed item difficutly) and the other is based on a latent trait model (Rasch difficulty). The third approach (chi-square) examined the distribution of correct and incorrect responses across different ability levels. (See Berk, 1982)

The transformed item difficulty or delta plot methods is based on standardizing the item difficulties for each group, performing a monotonic transformation (Angoff and Ford, 1973),

plotting the difficulties for each group on a two dimensional graph. If item difficulties are similar for the two groups, they will distribute themselves in the form of an ellipse around a 45^{o} regression line. If items do not behave similarly for the groups, they will fall at a distance from the regression line. The distance that these items are located from the regression line is an indicator of the degree of bias. In traditional studies of bias, these outliers would be termed 'biased items.' A series of SPSS programs was written and delta values plotted using a Versatec plotter.

The chi-square approach was devised by Scheuneman (1979). An item is defined as unbiased if all persons of the same ability, regardless of group membership, have an equal probability of a correct response. This procedure involves dividing the total test score range into a number of intervals (five were used in the study) and using the proportion of individuals within each range as an estimate of the probability of a correct response. A chi-square value contrasts the observed frequency versus expected frequency within each ability level across language groups. A large chi-square value would indicate that an item behaves differently for the two groups. A modified chi-square full procedure (Ironson, 1982) was used in this analysis.

The final parameter used Rasch latent trait item difficulty. Latent trait theory is based on the assumptions that a more able person has a better chance of success on an item than a less able person, and a person has a better chance of success on an easy item than a difficult one (Wright and Mead, 1977).

The item difficulty measure is often referred to as objective free because it provides a measure of item difficulty independent of the sample. Item difficulties were calculated in logits using the BICAL computer program developed at the University of Chicago and modified for use at Indiana University. These logits were then compared by using a statistic proposed by Draba (1974).

R E S U L T S

Table 1 presents the means and standard deviations for each TOEFL section by language. While the Arabic speakers scored higher on Section I: Listening, the Chinese speakers scored significantly higher on the other two sections.

Table 2 presents a comparison of reliabilities of TOEFL by section and language. All reliabilities were above .85 except for Chinese speakers on Section II which was .79. The difference between the reliabilities for the two native language groups in this section was .07. While the estimates of internal consistency were higher for Arabic native speakers in every section, there was no significant difference between the two groups as determined by Feldt F. Thus, there appear to be no significant differences in the accuracy of test measurement across groups for any test section.

Table 3 presents information on the factor similarity and percentage of variance accounted for by each test section for each language. The percentage of variance accounted for indicates the total amount of the total test score variance accounted for by that factor. The higher the variance, the more that factor explains of the test score. In every case the first factor

accounted for the largest proportion (at least 60 percent) of the variance; thus the degree of similarity for the first factor on each test section is most important. The coefficient of congruence for the first factor on each test section was .96. This indicates almost perfect agreement on the first factor for each test section. While the coefficients of congruence for other factor range from -.12 to .60, their importance is diminished by the low percentage of variance accounted for by these subsequent factors.

In summary, the results of the macro analysis indicate that all TOEFL sections are measuring similar constructs with the same degree of accuracy.

The micro analysis used three different techniques. Numerical values were calculated for each item by each method. Table 4 presents the range of values obtained by each bias technique. It must be noted that each method has its own scale of values and that a positive transformed difficulty value indicated an item was more difficult for the Arabic group and a positive Rasch value indicated a more difficult item for the Chinese group. Table 5 shows the continuum of the degree of bias detected by each method. Regardless of method, the micro test level was shown to be sensitive to native language background. Thus, at the micro level, evidence has been presented to show that items are sensitive to native language differences.

C O N C L U S I O N S

The purpose of this study was to provide an additional perspective for the construct validation of second language tests. A theoretical model from psychometric bias studies was modified and applied to examine the performance of different native language groups on tests of second language proficiency. This perspective was then applied to examine the performance of Arabic and Chinese speakers on the TOEFL.

While this study was limited to only one test and only two groups, it does suggest an additional perspective and methodology in the construct validation process. Of course by itself it does not establish construct validity, but combined with other methods and approaches it adds information to test construct validation.

This study also provides additional evidence for the construct validity of the TOEFL. The user should have more confidence that the test is measuring what it preports to measure.

One additional underlying conclusion/warning emanates from this study. Too often second language research and testing has suffered from what Oller (1980) has called "Psychometrosis", uneasiness of testing. In some cases, second langauge studies have accepted at face value the studies, results, and theories of statistical research design and psychometrics without properly understanding the nature of the theory. In other cases, psychometricians have conducted research without understanding the nature of second language training. For example, a strict psychometric interpertation of the data on differential performance by different

native language groups would have led to cries of "bias". While linguists concern is language theory and psychometricians measurement theory, studies in language testing need to be cognizant of the needs, demands, and theories of both. While theories of language learning and syntax may elude the psychometrician, proper methods of construct validation may elude the linguist. What is important for those concerned is to be aware and knowledgeable in both areas for mutual benefit. This study has attempted to blend the two in its examination of test construct validation process.

Table 1

Means and Standard Deviations of TOEFL by Section and Language

	Section	Language				
		Arabic[a]		Chinese[b]		
		Mean	Stand Dev	Mean	Stand Dev	
I	Listening Comprehension	30.38	9.1	29.49	7.8	
II	Structure and Written Expression	23.97	7.2	26.31	5.5	*
III	Reading Comprehension and Vocabulary	30.99	9.6	38.15	8.5	*

$P < .01$

Table 2

Comparison of Reliabilities of TOEFL by Section and Language

	Section	Number of Items	Language		Feldt F
			Arabic	Chinese	
I	Listening Comprehension	50	.83	.85	1.25n.s.
II	Structure and Written Expression	40	.76	.79	1.50n.s.
III	Reading Comprehension and Vocabulary	60	.88	.86	1.17n.s.

Table 3

Factor Similarities and Percentages of Variance accounted for by Section and Language

Section	Variance accounted for (%) Arabic	Chinese	Coefficient of Congruence
Section I: Listening			
Factor 1	73	71	.96
Factor 2	13	12	.50
Factor 3	7	9	.31
Factor 4	6	8	.22
Section II: Structure/Written			
Factor 1	69	70	.96
Factor 2	14	12	.60
Factor 3	10	10	.23
Factor 4	8	8	.17
Section III: Reading/Vocabulary			
Factor 1	71	63	.96
Factor 2	13	15	-.12
Factor 3	8	13	.06
Factor 4	7	9	.19

Table 4

Means, Standard Deviations, and Ranges of Different Item Bias Techniques by Section

Statistic	Transformed Difficulty	Chi-square	Rasch Difficulty
Section I: Listening Comprehension (50 items)			
Mean	.00	18.56	.04
Standard deviation	1.87	20.42	3.58
Minimum value	-4.40	.85	-6.93
Maximum value	3.74	103.62	8.95
Section II: Structure and Written Expression (40 items)			
Mean	.00	17.65	.11
Standard deviation	1.32	18.22	3.45
Minimum value	-3.56	2.01	-7.96
Maximum value	3.04	68.73	7.52
Section III: Reading Comprehension and Vocabulary (60 items)			
Mean	.00	12.94	-.07
Standard deviation	1.25	9.54	2.78
Minimum value	-2.96	2.36	-7.54
Maximum value	3.20	46.73	6.54

References

Alderman, D.L., & Holland, D. W. Item performance across native language groups on the Test of English as a Foreign Language (TOEFL Research Report 9). Princeton, N. J.: Educational Testing Service, 1981.

Bachman, L. F., & Palmer, A. The construct validation of the FSI oral interview. Language Learning, 1981, 31, 67-86.

Berk, R. A. Handbook of methods for detecting test bias. Baltimore: The Johns Hopkins University Press, 1982.

Campbell, D. T., & Fiske, D.W. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 1959, 56, 81-105.

Draba, R. E. The identification and interpretation of item bias. Research Memorandum No. 26. Statistical Laboratory, Department of Education, University of Chicago, 1977.

Harman, H. H. Modern factor analysis (2nd Ed.). Chicago: The University of Chicago Press, 1976.

Ironson, G. H. Use of chi-square and latent trait approaches for detecting item bias. In R. A. Berk (Ed.), Handbook of methods for detecting test bias. Baltimore: The Johns Hopkins University Press, 1982, 117-160.

Ironson, G. H., & Subovíak, M. J. A comparison of several methods of assessing item bias. Journal of Educational Measurement, 1979, 16, 209-225.

Jensen, A. R. Bias in mental testing. New York, N. Y.: The Free Press, 1980.

Oller, J. W. Evaluation and testing in vernacular language. Paper presented at the International Symposium on Standard Language/Vernacular Relations in Bilingual Education, Racine, Wisconsin, November 16-19, 1980. Proceedings to be edited and published by A. Valdman and B. Hartford).

Oller, J. W. & Hinofotis, F. B. Two mutually exclusive hypotheses about second language ability: Indivisible or partially divisible competence. In J. W. Oller and K. Perkins (Eds.), Research in language testing. Rowley, Mass.: Newbury House, 1980, 13-24.

Reynolds, C. R. Methods for detecting construct and predictive bias. In R. A. Berk (Ed.), Handbook of methods for detecting test bias. Baltimore: The Johns Hopkins University Press, 1982a, 199-227.

Scheuneman, J. A method of assessing bias in test items. Journal of Educational Measurement, 1979, 16, 143-152.

Shepherd, L. A., Camilli, G., & Averill, M. Comparison of procedures for detecting test-item bias with both internal and external ability criteria. Journal of Educational Statistics, 1981, 6, 317-372.

Stevenson, D. Construct validity and the test of English as a foreign language. Unpublished doctoral dissertation, University of New Mexico, 1974.

Wright, B. D. & Mead, R. J. BICAL: Calibrating items and scales with the Rasch model (Research Memorandum, Number 23). Chicago: University of Chicago, 1976.